Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W321–W326***
DOI: 10.1093/nar/gkh377

The PredictProtein server

Burkhard Rost[1,2,3,*], Guy Yachdav[1,3] and Jinfeng Liu[1,2]

[1]CUBIC and [2]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA and [3]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Received February 13, 2004; Revised and Accepted March 15, 2004

ABSTRACT

PredictProtein (http://www.predictprotein.org) is an Internet service for sequence analysis and the prediction of protein structure and function. Users submit protein sequences or alignments; PredictProtein returns multiple sequence alignments, PROSITE sequence motifs, low-complexity regions (SEG), nuclear localization signals, regions lacking regular structure (NORS) and predictions of secondary structure, solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide-bonds, sub-cellular localization and functional annotations. Upon request fold recognition by prediction-based threading, CHOP domain assignments, predictions of transmembrane strands and inter-residue contacts are also available. For all services, users can submit their query either by electronic mail or interactively via the World Wide Web.

OVERVIEW

PredictProtein (PP) is an automatic service that searches up-to-date public sequence databases, creates alignments, and predicts aspects of protein structure and function. Users send a protein sequence and receive a single file with results from database comparisons and prediction methods. PP went online in 1992 at the European Molecular Biology Laboratory (EMBL, Heidelberg); since 1999 it has operated from Columbia University (New York). Although many servers have implemented particular aspects, PP remains the most widely used public server for structure prediction: over 1.5 million requests from users in 104 countries have been handled; over 13 000 users submitted 10 or more different queries. PP web pages are mirrored in 17 countries on 4 continents. Our goal has always been to develop a system optimized to meet the demands of experimentalists not experienced in bioinformatics. This implied that we focused on incorporating only high-quality methods, and tried to collate results omitting less reliable or less important ones.

Attempt to simplify output by incorporating a hierarchy of thresholds

The attempt to 'pre-digest' as much information as possible to simplify the ease of interpreting the results is another unique pillar of PP. For example, by default PP returns only those proteins found in the database that are very likely to have a similar structure to the query protein (1). Particular predictions, such as those for membrane helices, coiled-coil regions, signal peptides and nuclear localization signals, are not returned if found to be below given probability thresholds. Over the years, we have added so many methods into the output of PP that our original 'easy-to-interpret' goal is challenged. We hope that a variety of improvements in the near future will reduce this problem.

Each request triggers the application of over 20 different methods

Currently, users receive a single output file with the following results (some of these are optional, Table 1). Database searches: similar sequences are reported and aligned by a standard, pairwise BLAST (2), an iterated PSI-BLAST search (3) and by the dynamic-programming method MaxHom (4). Although the pairwise BLAST searches are identical to those obtainable from the NCBI site, the iterated PSI-BLAST is performed on a carefully filtered database to avoid accumulating false positives during the iteration (5,6). The dynamic-programming method MaxHom is only available through PP. In addition, database searches comprise a standard BLAST-based search through ProDom (7) and a standard search for functional motifs in the PROSITE database (8). PP now also identifies putative boundaries for structural domains through the CHOP procedure (below). Optionally, users can request searches for remotely similar proteins by the prediction-based threading method TOPITS (9,10). Structure prediction methods: secondary structure, solvent accessibility and membrane helices predicted by the PHD and PROF programs (11,12, B. Rost, manuscript submitted), membrane strands predicted by PROFtmb (H. Bigelow, D. Petrey, J. Liu, D. Przybylski and B. Rost, manuscript submitted), coiled-coil regions by COILS (13), bonded cysteine residues by

*To whom correspondence should be addressed. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: rost@columbia.edu

Table 1. Methods used by PP

Method	Task	Main author(s)	Quote
Database			
Swiss-Prot[a]	Annotated protein sequences	A. Bairoch (SIB) and R. Appweiler (EBI)	(44)
TrEMBL[a]	Raw protein sequences	R. Appweiler (EBI)	(44)
PDB[a]	Protein structures	P. Bourne (UCSD)	(45)
BIG	Non-redundant combination of Swiss-Prot, TrEMBL, PDB	D. Przybylski (Columbia)	(5)
Alignment			
MaxHom	Dynamic programming, multiple alignment	R. Schneider (LION) and C. Sander (Sloan Kettering)	(4)
BLASTP[a]	Pairwise alignment	S. Karlin and S. F. Altschul (NCBI)	(2)
PSI-BLAST[a]	Profile based alignment	S. F. Altschul (NCBI)	(3)
HMMer[a]	Hidden Markov model search	S. Eddy (Washington University)	(36)
TOPITS	Prediction-based threading	B. Rost	(9,46,47)
Protein domains and unusual regions			
ProDom[a]	Structural domain-like regions	F. Corpet, F. Servant, J. Gouzy and D. Kahn (Toulouse)	(48)
Pfam-A[a]	Protein families	A. Bateman (Sanger) *et al.*	(35)
CHOP	Structural domain-like fragments	J. Liu (Columbia)	(33)
SEG[a]	Low-complexity regions	J. C. Wootton and S. Federhen (NCBI)	(18)
NORSp	Floppy regions	J. Liu and B. Rost	(19,20)
Protein structure			
PHDsec	Secondary structure	B. Rost	(11,49,50)
PHDacc	Solvent accessibility	B. Rost	(11,51)
PHDhtm	Membrane helices	B. Rost	(11,52,53)
PROFsec	Secondary structure	B. Rost	(12)
PROFacc	Solvent accessibility	B. Rost	Unpublished
GLOBE	Globularity	B. Rost	Unpublished
COILS	Coiled-coil regions	A. Lupas (Tübingen)	(54)
CYSPRED[a]	Disulphide bonds	P. Fariselli and R. Casadio (Bologna)	(14)
ASP	Structural switches	M. Young and S. Highsmith (Sandia)	(17)
PROFcon08	Inter-residue contacts	M. Punta (Columbia)	(15)
PROFtmb	Membrane barrels	H. Bigelow	(Manuscript submitted)
Protein function			
PredictNLS	Nuclear localization signals	R. Nair, M. Cokol and B. Rost (Columbia)	(21,22)
PROSITE[a]	Functional sequence motifs	K. Hofmann, P. Bucher and A. Bairoch (SIB)	(8)
LOCnet	Prediction of sub-cellular localization	R. Nair	(24)
Tools integrated into PP			
MView[a]	HTML alignment viewer	N. Brown	(55)
ESPript[a]	Ready-to-publish alignments and predictions	P. Gouet and E. Courcelle (IPS Toulouse)	(56)

[*]Original URLs: Swiss-Prot, http://www.expasy.org/sprot/; TrEMBL, http://www.ebi.ac.uk/trembl/; PDB, http://www.rcsb.org/pdb/; BLASTP/PSI-BLAST, http://www.ncbi.nlm.nih.gov/BLAST/; HMMer, http://hmmer.wustl.edu/; ProDom, http://protein.toulouse.inra.fr/prodom.html; Pfam-A, http://www.sanger.ac.uk/Software/Pfam/; SEG, http://trex.musc.edu/manuals/unix/seg.html; CYSPRED, http://prion.biocomp.unibo.it/cyspred.html; PROSITE, http://www.expasy.org/prosite/; Mview, http://mathbio.nimr.mrc.ac.uk/~nbrown/mview/; ESPript, http://prodes.toulouse.inra.fr/ESPript.

CYSPRED (14) and inter-residue contacts through PROFcon08 (15). Putative structural switching regions are detected by the program ASP (16,17), low-complexity regions are marked by SEG (18) and long regions with no regular secondary structure are identified by NORSp (19,20). The PHD/PROF programs and TOPITS are only available through PP. The particular way in which PP automatically iterates PSI-BLAST searches and the way in which we decide what to include in sequence families is also unique to PP. The particular aspects of function that are currently embedded explicitly in PP are all somehow related to sub-cellular localization: we detect nuclear localization signals through PredictNLS (21,22) and endoplasmic reticulum and Golgi-related signals through another in-house data set (23); we predict localization independent of targeting signals through LOCnet (24); and we annotate homology to proteins involved in cell-cycle control (25).

PERFORMANCE OF METHODS

A detailed review of the strengths, weaknesses and pitfalls of the many methods applied by PP is far beyond the scope of this description. We give only a brief overview of trends in the following.

(i) *Alignment methods*: While the dynamic-programming method MaxHom still appears best in aligning pairs of proteins, the iterated PSI-BLAST tends to be more sensitive in unravelling more distantly related proteins and also in correctly aligning them provided the underlying profiles contain enough information. Note, however, that PSI-BLAST tends to over-estimate the relevance of short matches and that PSI-BLAST expectation values have to be viewed with extreme caution when inferring similarity in function (26–28).

(ii) *Protein domains and unusual regions*: Like for instance SMART (29), ProDom tends to identify regions that are significantly shorter than structural domains (30); this is not the case for CHOP. However, CHOP misses many domain boundaries since it relies heavily on similarities to domains annotated by others (PrISM, Pfam-A). Note that short regions of low-complexity (SEG) are fairly common and are not necessarily informative.

(iii) *Protein structure [EVA (31)] for an up-to-date evaluation of structure prediction]*: (a) PROFsec secondary structure prediction: on average, 76% of all residues are

correctly predicted (only ~71% by PHDsec). (b) PROFacc accessibility prediction: almost 80% of all residues are correctly predicted as either buried or exposed, and >80% of the surface residues are correct. (c) PHDhtm membrane helix prediction: ~80% of the membrane helices are correctly predicted; for ~66% of all tested proteins, all membrane helices and the topology were correctly predicted (32); at the default threshold, membrane helices are incorrectly detected in ~2% of the tested globular proteins (32); about one-quarter of all signal peptides (for secreted proteins) are mistaken for membrane helices (32). (d) PROFtmb membrane barrel prediction: at high levels of reliability PROFtmb never confuses proteins with and without membrane strands; >80% of the membrane strand residues are correctly predicted; up- and down-strands are rarely confused. (e) GLOBE: not accurate enough to identify domain boundaries. GLOBE often correctly captures trends such as 'very unlike a globular protein'. Multi-domain proteins with globular and non-globular domains—such as NORS regions—are misclassified, however. (f) COILS: perceived to be correct most of the time. (g) CYSPRED: most disulfide-bonding residues are correctly identified; however, most predicted bonds are wrong. (h) ASP: if the protein has a structural switching region, this is usually detected correctly. (i) PROFcon08: most inter-residue contacts that are predicted are wrong; in fact, even at a coverage of 10%, only 27–40% of all contacts are correctly predicted. (j) NORSp predictions of non-regular regions: so far, there is no example of a protein with regular structure that we predicted to be irregular. Note that the PROF and PHD series and CYSPRED are all based on the artificial neural network systems (except for PROFtmb, which is based on a hidden Markov model).

(iv) *Protein function*: our signal-motif-based predictions reach levels of accuracy from as high as close to 100% (NLS) to as low as 50% (endoplasmic reticulum and Golgi apparatus). Homology transfer and keyword-based annotations are returned at levels >70% accuracy. Our system for *de novo* prediction of sub-cellular localization reaches levels ~60% accuracy (extra-cellular space, cytoplasm, nucleus, mitochondria, other).

NOVEL METHODS

CHOP (33) is a hierarchical procedure that chops proteins into structural domain-like fragments through similarity to domains of known structure [taken from PrISM (34)], or to Pfam-A domain-like fragments (35) [searches through HMMer (36)], or to full-length natively expressed proteins taken from Swiss-Prot (37). The major mistakes of CHOP result from incorrect original annotations (in PrISM or Pfam-A). The major shortcoming is that the procedure misses many domains that have no significant level of sequence similarity to known domain-like fragments. CHOP is currently an option, i.e. not run by default.

PROFtmb predicts beta-barrel membrane proteins, their topology and the residues in membrane strands (in four states). The method is so accurate in distinguishing proteins with and without beta-membrane barrels that at the default threshold we do not expect any error (H. Bigelow, D. Petrey, J. Liu, D. Przybylski and B. Rost, manuscript submitted). Over 80% of the residues are classified correctly into one of the four states up- and down-strand, inner- and outer-loop. PROFtmb is currently not run by default.

PROFcon08 appears to be one of the most accurate existing methods in predicting inter-residue contacts (15). However, this comes with a caveat: most non-local contacts predicted are not observed, and most observed contacts are not predicted. As a rule of thumb, if we predict one-tenth of the observed contacts, one-third of our predictions are right. PROFcon08 is currently not run by default.

We built a database of proteins involved in cell-cycle control [CellCycleDB (25)]. We used this database to estimate problem-specific levels of accuracy and coverage in homology-transferofexperimentalinformation.Theseestimates allow a controlled, automatic search with proteins against CellCycleDB. This search is currently not run by default.

METHODS TO BE ADDED BY SUMMER 2004

LOCnet appears to be the most accurate general method for the *de novo* prediction of sub-cellular localization with a four-state accuracy ~65% (24). Performance is best for extra-cellular and worst for mitochondrial proteins. LOCnet is currently not run by default.

CHOPnet is a neural network-based method for the *de novo* prediction of structural domains in fragments that could not be treated by CHOP (J. Liu and B. Rost, manuscript submitted). The method correctly predicts ~55% of all known two-domain proteins to have two domains; for about one-half of these the domain boundary is correctly placed within 20 residues of the observed boundary. Performance is worse for proteins with more than two domains. However, by pre-digesting the query with CHOP, in many cases the task for CHOPnet will resemble the prediction of single- or two-domain proteins (for which the prediction accuracy is reasonably high).

ISIS is a method that specifically predicts residues involved in transient, external protein–protein interactions (38,39). The current system is based on neural networks that use information from alignments and other prediction methods. The method returns predictions at different levels of accuracy/coverage: at 5% coverage the accuracy reaches about ~60%.

LOCi is a hierarchical system that predicts sub-cellular localization through a variety of sources, namely through homology to proteins of experimentally known localization [LOChom (27,40)], through Swiss-Prot keyword searches [LOCkey (41)], localization signals [SignalP (42)], TargetP (43), PredictNLS (21,22)] and a combination of *de novo* prediction methods based on support vector machines and neural networks (R. Nair and B. Rost, unpublished data). Prediction accuracy exceeds 70%, making the method the most comprehensive and most accurate means of predicting sub-cellular localization.

INPUT, OUTPUT AND JOB OPTIONS

Default output

The output format is self-documenting. The output contains

(i) a list of likely homologues found in the protein database (BIG) and—upon request—the multiple alignments of these sequences (by default in 'HTML' format from MView). Note that we have now switched to no longer

returning the entire PSI-BLAST alignments by default since these are often of considerable size.
(ii) If found, a list of the putative PROSITE motifs.
(iii) If found, a list of ProDom and/or CHOP domain-like fragment assignments.
(iv) If found, a prediction of coiled-coil regions.
(v) Information about the expected levels of accuracy of structure predictions.
(vi) Prediction of aspects of protein structure. These are grouped in the following way: (a) prediction of secondary structure for all residues; (b) prediction of secondary structure for reliably scored residues only, with an expected three-state accuracy for these residues of >85%; (c) prediction of solvent accessibility for all residues; (d) prediction of solvent accessibility for reliably scored residues only, with an expected correlation between experimental observation and prediction of 0.69; and (e) prediction of transmembrane helices and their topology (if any detected). Note that for the prediction of transmembrane helices and strands a conservative threshold is chosen. Thus, a membrane segment may not be detected using the default parameter settings.

Advanced input options

By default, users submit a protein through its one-letter residue sequence. However, PP also accepts submissions in FASTA, PIR and Swiss-Prot formats or through the Swiss-Prot identifier. Most prediction methods applied use the information from the multiple alignments created by PP; prediction accuracy increases with the quality of the alignment. PP's alignments are fully automated, thus may not be as accurate as an alignment that experts have hand-edited. Therefore, users may also submit their favourite alignment directly. PP accepts alignments as FASTA lists, PIR lists, as well as in SAF and MSF formats. The fold recognition/prediction-based threading method TOPITS uses predictions of secondary structure and solvent accessibility to search through a library of proteins of known structure. Predictions can be submitted through a simple column-based format.

Advanced prediction/job options

Not all methods are executed by default; some methods (such as the prediction of membrane helices) use particular 'conservative' thresholds when included automatically and different thresholds when requested explicitly. In particular, the following methods can be toggled (switched on or off): MaxHom, BLASTP, PSI-BLAST, SEG, PHDsec, PHDacc, PHDhtm, PROFsec, PROFacc, COILS, CYSPRED, ASP, PROSITE, ProDom, CHOP, NORSp, PROFtmb, PROFcon08, LOCkey, LOChom, PredictNLS and LOCnet. Users can also explicitly request TOPITS+ or can evaluate the prediction accuracy of a secondary structure prediction method (EvalSec). Note that switching off methods has two advantages: it speeds up the execution and it reduces the size of the output. However, bear in mind that the database searches and their results are the limiting factor for speed and bytes produced.

Advanced output options

The default output now is an HTML-formatted file, i.e. ready to display in any browser. Users can change this default to output in raw text in the following alignment formats: BLAST, no alignment, HSSP, HSSP profiles only, MSF, SAF and FASTA list. The results from the predictions are also available in a variety of machine-readable formats. (Developers: please do not write parsers for the human-readable PP output; if in doubt, contact us, since we can write almost any reasonable format if need be.) Due to the size of multiple alignments, we no longer email the results; rather the output will be stored for a week on our website (remember to download it in that period). Upon request, results are returned by email.

Interactive versus batch jobs

By default, the user submits requests to a batch queue and will be notified by email where to find the results (or will be sent these results). While PP also has an interactive mode that will write the results directly into the requesting web browser, this option comes with a restriction on the length of time for which the web connection is kept open: if PP has not completed a request within 5 min, we automatically switch the job to a batch mode and notify users by email. In practice, this implies that interactive jobs will only finish in time if (i) the PP queue is empty (works on a first-come-first-served principle) and (ii) the request does not require more than 5 min of CPU time (typically the case if an alignment is submitted, and/or the query protein is short and/or has few homologues in today's databases). We have just upgraded the CPU resources for PP (now running on a LINUX farm); this has increased the probability of successful interactive queries.

Job queuing system

In order to maximize processor usage, requests to PP are queued and maintained by a mechanism that balances the work load by monitoring the status of the 10 CPUs currently dedicated to the server in normal operation. Users can query job and overall workload statuses through the web interface.

Portable versions

Most in-house programs are—or will be—available under general GNU licences (free for academia). Porting the entire PP system is a more complicated enterprise. We are currently optimizing the system to increase its portability. It is now available for local LINUX and IRIX installations. Furthermore, to make the system less bound to local OS and hardware constraints, future plans include decoupling some of the core services from the rest of the system and handling communication using innovative technologies such as XML-RPC or SOAP.

ACKNOWLEDGEMENTS

Making PredictProtein survive a decade was a major effort; many colleagues helped with hands and brains; thanks to all of them! For the first years at EMBL, thanks to Antoine de Daruvar (Bordeaux University), Reinhard Schneider (EMBL, Heidelberg), Sean O'Donoghue (LION Biosciences, Heidelberg) and Chris Sander (Sloan Kettering, New York). Thanks to Rolf Appweiler for his continued support at the European Bioinformatics Institute (EBI-EMBL, Hinxton, England), and to Volker Eyrich (Schrödinger, New York) for software support during the move to the USA. Further

thanks to all who set up mirror pages and who consented to our using their software, in particular to Nigel Brown for MView, to Emmanuel Courcelle and Patrice Gouet (IPBS, Toulouse) for ESPript, to Florencio Pazos (London) for Threadlize, to Andrei Lupas (Max Planck, Tübingen) for COILS, to Piero Fariselli and Rita Casadio (Bologna University) for CYSPRED, to Reinhard Schneider (EMBL, Heidelberg) for MaxHom, to Malin Young (Sandia Labs, Albuquerque) for ASP, and to Rajesh Nair (Columbia University) for his methods predicting sub-cellular localization, and to Dariusz Przybylski (Columbia) for his invaluable scripts optimizing automatic PSI-BLAST searches. Last, not least, thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Cathy Wu (PIR/PSD), Phil Bourne (San Diego University), and their crews for maintaining excellent databases and to all experimentalists who enable computational biology!

PredictProtein has attracted its first public support from grant R01 LM07329-01 from the National Library of Medicine.

REFERENCES

1. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
2. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
3. Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
5. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 195–205.
6. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
7. Corpet,F., Gouzy,J. and Kahn,D. (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic. Acids Res.*, **27**, 263–267.
8. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
9. Rost,B. (1995) TOPITS: Threading One-dimensional Predictions into Three-dimensional Structures. In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Third International Conference on Intelligent Systems for Molecular Biology*, Cambridge, England. AAAI Press, Menlo Park, CA, pp. 314–321.
10. Przybylski,D. and Rost,B. (2004) Improving fold recognition without folds. *J. Mol. Biol.* in press.
11. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525–539.
12. Rost,B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
13. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
14. Fariselli,P., Riccobelli,P. and Casadio,R. (1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins*, **36**, 340–346.
15. Punta,M. and Rost,B. (2004) Toward good 2D predictions in proteins. *FEBS* in press.
16. Kirshenbaum,K., Young,M. and Highsmith,S. (1999) Predicting allosteric switches in myosins. *Protein Sci.*, **8**, 1806–1815.
17. Young,M., Kirshenbaum,K., Dill,K.A. and Highsmith,S. (1999) Predicting conformational switches in proteins. *Protein Sci.*, **8**, 1752–1764.
18. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
19. Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
20. Liu,J. and Rost,B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, **31**, 3833–3835.
21. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localisation signals. *EMBO Rep.*, **1**, 411–415.
22. Nair,R., Carter,P. and Rost,B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res.*, **31**, 397–399.
23. Wrzeszczynski,K.O. and Rost,B. (2004) Annotating proteins from Endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. *CMLS*, in press.
24. Nair,R. and Rost,B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, **53**, 917–930.
25. Wrzeszczynski,K.O. and Rost,B. (2004) Cataloguing proteins in cell cycle control. *Methods Mol. Biol.*, **241**, 219–233.
26. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
27. Nair,R. and Rost,B. (2002) Sequence conserved for sub-cellular localization. *Protein Sci.*, **11**, 2836–2847.
28. Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
29. Ponting,C.P., Schultz,J., Milpetz,F. and Bork,P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.
30. Liu,J. and Rost,B. (2003) Domains, motifs, and clusters in the protein universe. *Curr. Opin. Chem. Biol.*, **7**, 5–11.
31. Koh,I.Y.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Narayanan,E., Graña,O., Valencia,A., Sali,A. and Rost,B. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.
32. Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.
33. Liu,J. and Rost,B. (2004) CHOP proteins into structural domains. *Proteins* (in press).
34. Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–711.
35. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
36. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
37. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
38. Ofran,Y. and Rost,B. (2003) Predict protein–protein interaction sites from local sequence information. *FEBS Lett.*, **544**, 236–239.
39. Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, **325**, 377–387.
40. Nair,R. and Rost,B. (2003) LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Res.*, **31**, 3337–3340.
41. Nair,R. and Rost,B. (2002) Inferring sub-cellular localisation through automated lexical analysis. *Bioinformatics*, **18**, S78–S86.
42. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
43. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
44. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
45. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
46. Rost,B., Schneider,R. and Sander,C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, **270**, 471–480.
47. Rost,B. (1995) Fitting 1-D predictions into 3-D structures. In Bohr,H. and Brunak,S. (eds.), *Protein Folds: A Distance Based Approach*. CRC Press, Boca Raton, FL, pp. 132–151.
48. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

49. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
50. Rost,B. and Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.
51. Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.
52. Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.*, **4**, 521–533.
53. Rost,B., Casadio,R. and Fariselli,P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.
54. Lupas,A. (1996) Prediction and analyis of coiled-coil structures. *Meth. Enzymol.*, **266**, 513–525.
55. Brown,N., Leroy,C. and Sander,C. (1998) MView: a Web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
56. Gouet,P., Courcelle,E., Stuart,D.I. and Metoz,F. (1999) ESPript: multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.